L'édition lexicographique dans un système générique de gestion de bases lexicales multilingues

Gilles Sérasset et Mathieu Mangeot-Lerebours
GETA-CLIPS IMAG, BP 53, 38041 Grenoble cedex 9, France
Gilles.Serasset@imag.fr – Mathieu.Mangeot@imag.fr

Résumé

Nous étudions les problèmes posés par la création d'outils pour lexicographes dans le cadre d'un système de gestion de bases lexicales multilingues génériques. Nous montrons une approche économique applicable dans de nombreux cas. Sa généralisation à des dictionnaires complexes n'étant pas possible, nous verrons les solutions envisageable suivant les situations rencontrées. Nous montrons les différents outils créés pour certaines de ces situations.

Mots clés

Traitement Automatique des Langues Naturelles ; base lexicale multilingue ; lexicographie ; dictionnairique.

1. Introduction

Dans le domaine de l'ingénierie linguistique et de la connaissance, le problème de la construction de ressources lexicales s'est toujours posé. Néanmoins, l'avancée des techniques du Traitement Automatique des Langues Naturelles (TALN) l'a rendu plus sensible. Il nous faut maintenant pouvoir répondre à des besoins importants en terme de quantité, de qualité et de complexité.

Pour simplifier la construction et l'évolution de ces nouveaux dictionnaires, (Sérasset 94b) propose un système générique de gestion de bases lexicales multilingues nommé SUBLIM (Système Universel pour les Bases Lexicales Multilingues). Ce système permet au lexicologue, grâce à un langage de haut niveau, de définir une base lexicale (un ensemble de dictionnaires) et de spécifier la forme des structures lexicales contenues dans chaque dictionnaire.

En échange de cet effort de formalisation, SUBLIM fournit des outils de vérification de cohérence (locale et globale), de défautage, et d'import/export.

L'avantage principal de SUBLIM par rapport à d'autres projets analogues est qu'il n'impose pas une structure informatique unique pour le codage de la microstructure des dictionnaires. Ainsi, le lexicologue a la possibilité de coder l'information qu'il désire sous la forme qu'il désire en utilisant la structure informatique la plus adaptée. Le système SUBLIM respecte ainsi le critère de **fidélité linguistique** énoncé par (Shieber 86).

Ce critère doit aussi être respecté dans la structure externe manipulée par le lexicographe. Le travail sur l'interface d'édition d'un dictionnaire défini en SUBLIM est donc lui aussi crucial.

Si le développement d'architectures logicielles avancées (OpenDoc par Apple™, ou les Java Beans par SUN™) nous apporte des solutions pour des bases lexicales complexes, il existe des solutions plus simples et plus efficaces qui peuvent être mises en œuvre pour une grande majorité de dictionnaires.

Nous discutons dans cet article des outils que nous avons utilisés pour la construction de trois dictionnaires très différents. Après avoir situé le contexte de ce travail, nous présenterons la construction de ces trois dictionnaires en motivant les choix effectués. Nous conclurons en donnant les perspectives que nous envisageons à la suite de ces expériences.

2. Contexte

2.1 Systèmes de gestion de bases lexicales

De nombreux projets se sont intéressés aux problèmes du lexique pour la linguistique informatique. Certains l'ont abordé sous l'angle de la standardisation des représentations du dictionnaire (Genelex 93). D'autres, comme le projet Multilex, proposent une structure linguistique commune pour différentes langues de la communauté européenne, tout en laissant au lexicologue la possibi-

lité de rajouter les informations qu'il désire, pourvu qu'il code ses informations dans le formalisme du système : les structures de traits typés. Ce choix d'une structure unique ayant été fait par soucis de simplicité.

Pourtant, en étudiant de près quelques dictionnaires, on trouve une très grande diversité dans les choix linguistiques (macro et microstructure) et dans les choix de structures informatiques. Ainsi, le projet Genelex se base sur une structure entité-relation que l'on peut interpréter comme un graphe. Les dictionnaires du LADL (Gross 87) utilisent des automates. D'autres, enfin, se basent sur des arbres, des fonctions, …

Nous défendons la thèse selon laquelle ce n'est pas au lexicologue d'adapter ses choix de représentation à l'outil utilisé, mais c'est à l'informaticien d'adapter ses outils aux différents besoins des lexicologues.

Ce credo est néanmoins difficile à mettre en œuvre dans un contexte où il n'y a pas d'informaticien. C'est précisément pour cela que nous avons créé le système SUBLIM.

2.2 SUBLIM

Pour utiliser le système SUBLIM, le lexicologue doit décrire la structure interne de sa base lexicale en utilisant deux langages de haut niveau (Sérasset 94a et 94b).

Le premier lui permet de définir l'architecture lexicale de sa base. Il va spécifier l'ensemble des dictionnaires de la base et leur type (monolingue, bilingue, interlingue). Il peut ainsi définir une base lexicale basée sur une approche par transfert ou sur une approche par pivot.

```
(def-linguistic-class french_entry
  (feature-structure
   (lexical_unit string)
   (Part-of-Speech
    (one-of "n.m" "n.f" "v.t" "v.i" "v.pr."
            "a" "adv" "loc" "prep"))
   (example (set-of string))
   (indexer string)
   (quality (one-of manual auto reviewed))
   (properties (set-of property))
   (uws (set-of string))
   ))
```

Figure 1 : la description d'une unité lexicale dans le langage de SUBLIM (cf figure 6, §4.2).

Il définit ensuite, pour chaque dictionnaire, les structures informatiques des unités de son lexique. Pour cela, il utilise les constructeurs de base du langage : arbre, graphe, automate, structure de traits, liste, ensemble, énumération, etc. Nous donnons dans la figure 1 la structure d'unité lexicale (décrite figure 5, §4.1) dans le langage de SUBLIM.

2.3 L'édition lexicale

Il nous reste à aborder le thème principal de cet article : *quels outils peut-on proposer aux lexicographes pour éditer le plus simplement possible les bases lexicales définies en SUBLIM ?*

Lorsqu'on parle d'une interface d'édition lexicale, les problèmes deviennent plus complexes. Ainsi la *forme* que l'on va donner aux informations ne dépendent pas seulement de leur structure interne, mais aussi de l'interprétation qu'a le lexicologue de ces structures.

De plus, on peut envisager différentes interfaces pour les différents lexicographes amenés à travailler sur la base.

Enfin, suivant le contexte, le processus d'édition peut s'effectuer par un ou plusieurs lexicographes, avec ou sans le contrôle d'un administrateur de la base, localement ou par des pigistes travaillant chez eux, etc…

Ces différents contextes déterminent fortement le choix des outils que l'on va utiliser. Afin d'illustrer notre propos, nous allons maintenant exposer les choix effectués pour l'édition de :

- un dictionnaire français -> malais à usage humain,
- un dictionnaire français <-> anglais désambiguïsé pour un projet de traduction automatique basé sur une approche interlingue,
- une version informatisée du dictionnaire explicatif et combinatoire du français contemporain (Mel'ãuk et al. 95).

3. Dictionnaire français malais

3.1 Présentation

Le dictionnaire français-malais "Kamus Perancis-Melayu Dewan" (Gut et al. 96), a été construit en coopération entre le service Culturel de l'Ambassade de France à Kuala Lumpur, le Dewan

Bahasa dan pustaka, l'Unit Terjemahan Melalui Komputer (Universiti Sains Malaysia, Penang) et le Groupe d'Étude pour la Traduction Automatique (Université Joseph Fourier, Grenoble).

Du fait de la difficulté de trouver suffisament de lexicographes compétents en français et malais, le travail a débuté sur la base d'un dictionnaire français-anglais. Les entrées (français-anglais)-malais étaient ensuite révisées par un lexicographe expérimenté.

Pour ce travail d'indexage, nous avons utilisé un outil du commerce : le traitement de texte Word. Ce choix présentait de nombreux avantages :

- il fonctionne sur Mac et sur PC,
- les lexicographes savaient déjà l'utiliser,
- il était déjà disponible sur les machines des partenaires,
- contrairement à de nombreux outils d'indexage, il permet de voir tout un ensemble d'entrées de manière compacte et d'utiliser le copier/coller.

Les entrées du dictionnaire se présentent, en Word sous forme d'un suite de paragraphes. Chaque paragraphe contient un élément d'information. Le style du paragraphe permet de savoir de quel élément il s'agit (figure 2).

Figure 2 : le dictionnaire français anglais malais, tel qu'il est manipulé par le lexicographe

Les lexicographes travaillent sur des fichiers RTF (Rich Text Format) que nous analysons et intégrons dans la base. Les outils construits pour cela (analyse du RTF) ont été très simples à créer,

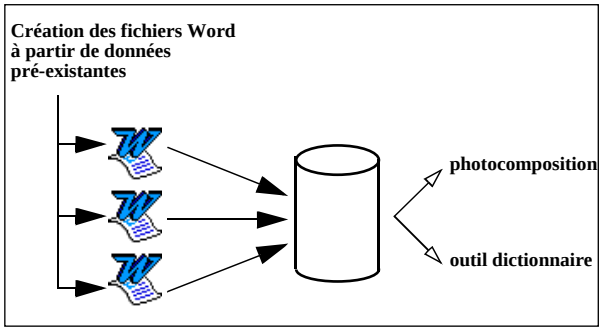

Figure 3 : Méthodologie de création du dictionnaire français malais.

du fait de l'utilisation d'un paragraphe par élément d'information. Nous avons même pu, dans un premier temps, utiliser les outils de Recherche/Remplacement intégrés à Word qui nous ont permis de créer, sans aucun effort, des fichiers texte balisés utilisables directement par la base centrale (Gaschler et Lafourcade 94). La méthodologie employée est schématisée dans la figure 3.

3.2 Bilan

L'utilisation d'un traitement de texte simple a permis la construction de ce dictionnaire de 20 000 entrées. Le principal avantage de cette méthode est sa simplicité. Les principaux développements informatiques ont porté sur l'exploitation de la base et non sur sa création. Le seul développement nécessaire pour la création de la base a été l'analyse des fichiers Word.

La distribution du travail entre les différents lexicographes est, elle aussi, très simple mais ce mode de distribution (basé sur l'échange des fichiers Word) est assez rigide.

Enfin, la création d'un dictionnaire est un processus assez long. Aussi, il est bon de compter sur un logiciel qui évolue durant tout ce temps. La contrepartie de cet avantage est que l'on est dépendant d'un format propriétaire. Néanmoins le sous ensemble du format RTF que nous utilisions a toujours été compatible avec les versions de Word utilisées.

L'inconvénient majeur de cette méthode est qu'il n'existait aucun outil permettant au lexicographe de vérifier le travail en cours. On ne peut, en effet, constater la malformation d'une entrée que lorsqu'on l'intègre à la base. Aussi, ce processus d'intégration ne peut se faire que sous le contrôle d'un administrateur lexicologue chargé de corriger

les erreurs des lexicographes (mauvais choix de style, abréviation inconnue, etc.).

4. Dictionnaire de traduction automatique

4.1. Présentation

UNL est un projet de communication multilingue interpersonnelle. L'aspect traduction automatique de ce projet se base l'UNL (Universal Networking Language, un langage pivot basé sur de l'anglais désambiguïsé). Chaque partenaire est en charge de la traduction de sa langue (pour nous, le français) vers l'UNL et vice versa. Pour construire notre système de traduction français <-> UNL, Il fallait rapidement construire une base bilingue à usage machinale. Nous sommes partis de différentes informations compilées à partir de dictionnaires disponibles, que nous avons intégrées dans notre base.

Pour vérifier et compléter la base, nous avons travaillé avec des lexicographes pigistes, étudiants en traduction. Nous avons donc repris la solution employée pour la construction du dictionnaire français malais (utilisation de fichiers Word, figure 4).

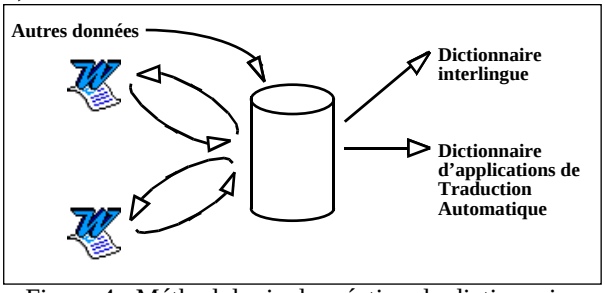

Figure 4 : Méthodologie de création du dictionnaire français <-> UNL.

Étant destinée à un usage machinal (un moteur de traduction automatique), certaines parties de cette information est codée. Or les lexicographes n'étaient pas formés pour ce type d'information. Aussi, seule une partie de cette information était fournie au lexicographe. Le reste était édité directement par un expert (figure 5).

À partir de la base (pré-remplie automatiquement) et d'une description des entrées sous forme de grammaire, nous générons automatiquement des fichiers RTF. Ces fichiers sont ensuite distribués aux lexicographes qui travaillent à domicile.

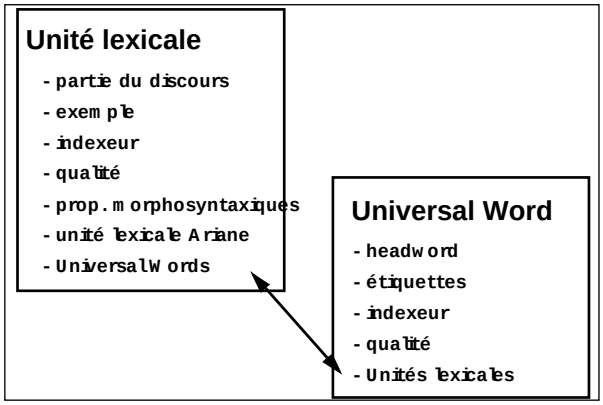

Figure 5 : structure interne de la base lexicale français <-> UNL. Seules les informations en gras sont données aux lexicographes

Le gestionnaire de la base récupère automatiquement les fichiers RTF et décide soit de les valider, soit de les renvoyer à l'indexage.

Enfin, en utilisant les macros de Word, nous avons créé quelques outils d'aide au lexicographe (vérification de la validité d'une entrée, liste des abréviations autorisées, guidage lors de l'utilisation des styles, …).

4.2. Bilan

La technique de construction du dictionnaire français-malais a été améliorée :

- nous pouvons générer les fichiers RTF à partir de la base existante,
- nous proposons des outils d'aide à l'indexage qui permettent au lexicographe de vérifier la structure d'une entrée.

La génération, les outils et la récupération sont paramétrables avec la structure des entrées du dictionnaire à condition que celle-ci reste relativement simple (descriptible par une grammaire LL1). Si l'ensemble du dictionnaire ne satisfait pas ce critère ou si certaines parties sont sensibles, les lexicographes peuvent tout de même travailler sur un sous ensemble de l'information disponible.

Nous avons donc généralisé cette méthode et possédons maintenant les outils qui nous permettront dans l'avenir de créer de nouveaux dictionnaires (Mangeot-Lerebours 98).

Il reste tout de même des problèmes inhérents à la méthode. Ainsi, même si nous fournissons au lexicographe des outils permettant de vérifier la structure des entrées, des problèmes subsistent lors de la récupération. En effet, ces outils ne fonction-

nent que lorsque le lexicographe les invoque. L'expérience montre qu'il ne le fait que rarement.

D'autre part, la présence d'un administrateur de la base est toujours nécessaire pour vérifier le travail des indexeurs.

5. Dictionnaire Explicatif et Combinatoire informatisé

Le projet NADIA-DEC (Sérasset 97), réalisé en collaboration entre le GETA et le GRESLET (département de linguistique et traduction de l'université de Montréal) a pour but l'informatisation du Dictionnaire Explicatif et Combinatoire du Français Contemporain (DEC). Au départ, le dictionnaire est disponible sous forme de fichier Word correspondant à la version imprimée (Mel'ãuk et al. 84, 88, 92).

L'approche utilisée ne remet pas en cause la structure linguistique que l'on peut trouver dans le DEC. La structure informatique du DEC doit permettre, au minimum, de re-générer à l'identique les fichiers Word utilisés pour la version papier. Aussi, toutes les informations sont présentes, et ce même si elles ne sont pas structurées. Il est ainsi toujours

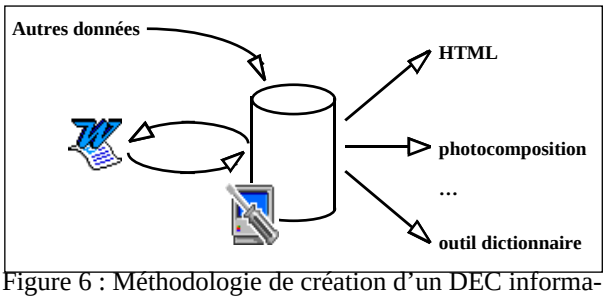

Figure 6 : Méthodologie de création d'un DEC informatisé.

possible, au fur et à mesure que l'on avance dans ce projet, d'augmenter la structuration des données sans avoir à reprendre l'ensemble du processus de récupération à partir des fichier Word.

À cet égard, le projet NADIA-DEC se distingue des autres projets d'informatisation du DEC qui se basent a priori sur une structure informatique simplifiée et qui n'informatisent que le sous ensemble de données commun entre le DEC et cette structure.

Le DEC est un dictionnaire très complexe. L'utilisation de Word comme interface pour lexicographe n'est donc pas possible (même si nous sommes parvenus par ailleurs à récupérer/regénérer les fichiers originaux). Nous avons donc créé un édi-

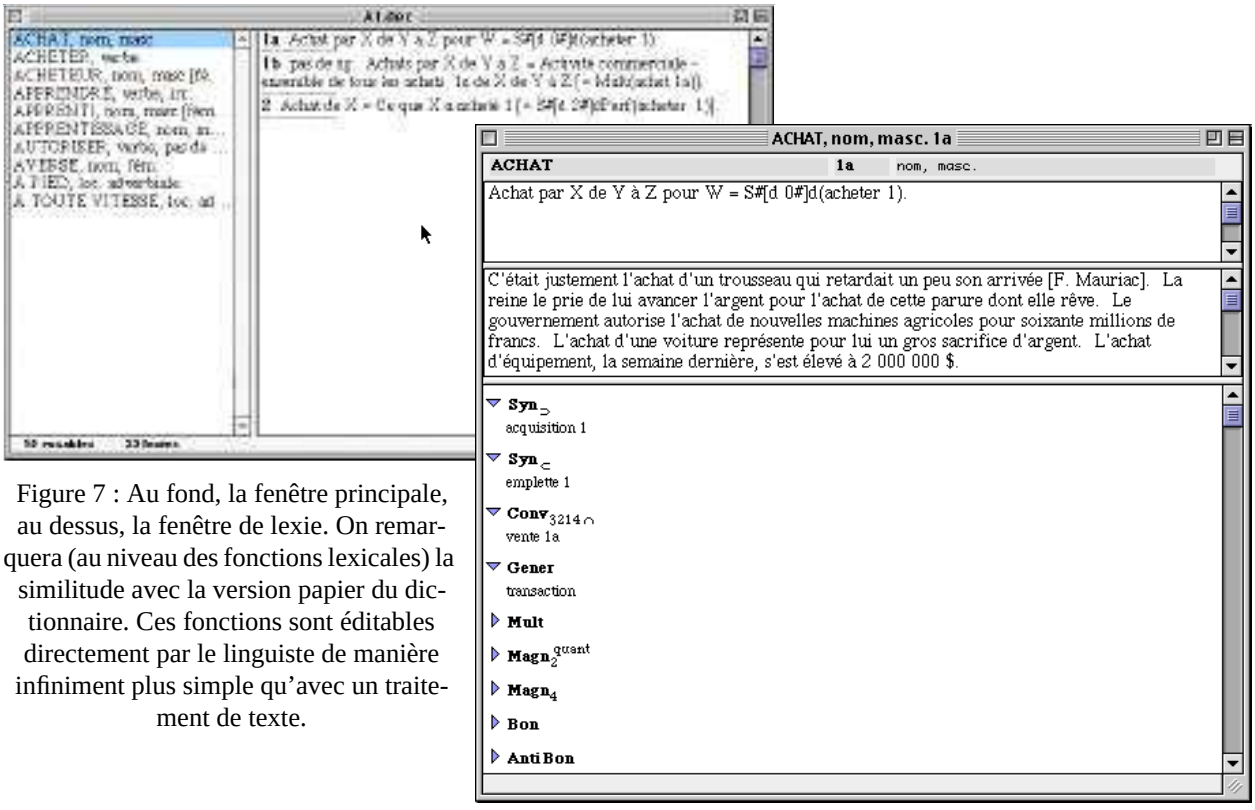

Figure 7 : Au fond, la fenêtre principale, au dessus, la fenêtre de lexie. On remarquera (au niveau des fonctions lexicales) la similitude avec la version papier du dictionnaire. Ces fonctions sont éditables directement par le linguiste de manière infiniment plus simple qu'avec un traitement de texte.

teur spécialisé du DEC : **DECID**. Dans la méthodologie adoptée (figure 6), l'édition se fait donc directement au niveau de la base lexicale.

Lors de la construction de DECID, nous avons mis l'accent sur le confort du lexicographe. Nous avons donc travaillé sur une interface directement inspirée de la version papier du DEC.

Le lexicographe dispose d'une fenêtre principale lui donnant la liste des vocables et des lexies du fichier en cours d'édition. La seconde fenêtre présente une lexie et permet de l'éditer (figure 7).

L'édition des fonctions lexicales est une tâche difficile lorsque les lexicographes travaillent sur un traitement de texte. Il faut faire attention à bien mettre les majuscules au bon endroit, passer en exposant ou en indice les parties qui doivent l'être, etc. Bref, au lieu de travailler sur la *signification* d'une fonction lexicale, le lexicographe travaille sur sa *forme*.

Avec DECID, le lexicographe peut éditer la fonction $\mathbf{Perm}_1\mathbf{IncepReal}_{3\subset}^{\text{usual}}$ simplement en tapant la séquence : `perm1incepreal3+'usual`. La mise en forme est totalement prise en charge par le logiciel.

5.2 Bilan

La structure informatique employée pour le dictionnaire DECID se base sur des types d'information assez simples (structures de traits, arbres, …). Par contre, son interface ne reflète pas directement cette structure, mais *l'interprétation* qu'en a le lexicologue. Ainsi, le lexicologue dispose d'un outil convivial et peut créer une entrée de manière très naturelle.

De plus, l'éditeur spécialisé ne permet pas de créer des entrées qui ne soient pas conformes à la structure utilisée. Il est aussi possible de vérifier des contraintes linguistiques sur les entrées éditées.

Par contre, la création d'une telle application est très lourde. De plus, la structure linguistique du DEC est en constante évolution et le logiciel peut difficilement évoluer en parallèle.

Enfin, pour que le logiciel évolue et pour que les bogues soient corrigés, il faut qu'un informaticien y travaille à temps plein. Le coût de cette solution est donc bien souvent prohibitif.

6. L'édition lexicale dans un système générique

En résumé, l'utilisation d'un traitement de texte nous apporte à moindre frais :

- la possibilité de distribuer simplement le travail à différents lexicographes
- un environnement multi plate-forme,
- des fonctionnalités intéressantes (copier/coller, rechercher/remplacer, vue simultanée de plusieurs entrées, etc.).

L'utilisation d'un outil spécialisé nous apporte :

- une interface spécialisée conviviale,
- une vérification constante de la bonne formation des entrées.

Aussi, nous pensons que ces deux méthodes (édition déportée par traitement de texte/édition intégrée par outil spécialisé) doivent être disponibles dans un outil générique. Elles pourront ainsi être utilisées conjointement dans des contextes différents (production de masse vs. vérification, sous-ensemble simple d'information vs. information complexe, lexicographe débutant vs. lexicographe expérimenté, …)

L'architecture du système SUBLIM est schématisée par la figure 8.

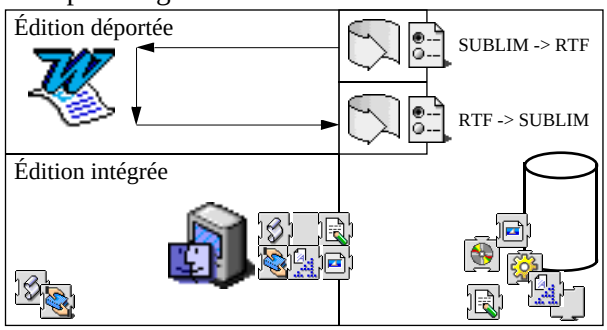

Figure 8 : Architecture de l'édition lexicale dans SUBLIM.

L'édition déportée est en majeure partie implémentée. La description de la structure informatique étant connue, il suffit d'attribuer un style à chaque élément d'information. On peut ainsi automatiquement générer les fichiers à compléter et les récupérer. Cette solution a déjà été utilisée pour le dictionnaire français-UNL.

L'édition intégrée reste à implémenter. La solution développée pourra être utilisé par un lexicologue sans intervention d'un informaticien. Ceci peut

être réalisé en s'inspirant de systèmes comme OpenDoc ou comme les java beans.

Le principe en est assez simple. L'utilisateur dispose d'un certain nombre de composants. Chaque composant sait afficher ou éditer un type d'information (arbre, graphe, structure de trait, chaîne de caractères…). En utilisant chacun de ces composants comme une brique de Lego, le lexicologue, après avoir défini sa structure linguistique, sera à même de définir son interface.

Cette solution est intéressante, mais n'est pas suffisante. En effet, la convivialité de l'éditeur DECID est due en grande partie au fait qu'il connaît non seulement la *structure* des données linguistiques, mais surtout qu'il en connaît *l'interprétation*. Il faut donc fournir aux utilisateurs de SUBLIM une API (Application Programming Interface), qui permettra à un informaticien de créer un composant spécialisé.

7. Conclusion

De toutes les fonctionnalités que doit fournir un système de gestion de base lexicales multilingues, l'édition lexicale influe directement sur la productivité des lexicographes. Dans un domaine où le besoin en lexique se fait de plus en plus ressentir, il faut donc accorder une attention particulière à cette fonctionnalité.

Néanmoins, dans notre travail, cet aspect vient en second. Nous considérons en effet qu'il est crucial de fournir un système ouvert permettant à tout lexicologue de définir, d'expérimenter, et de faire évoluer ses dictionnaire.

Nous pensons que les solutions évoquées sont un bon compromis entre ouverture et productivité.

Bibliographie

Gaschler J. et Lafourcade M. (1994). *Manipulating human-oriented dictionaries with very simple tools*, COLING-94, August 5-9 1994, vol. 1/2 : pp. 283-286.

Genelex (1993). *Projet Eureka Genelex, modèle sémantique*, Projet Eureka Genelex, Rapport Technique, 4 mars 1994, 185 p.

Gross M. (1987). *The Use of Finite Automata in the Lexical Representation of Natural Language*, Electronic Dictionaries and Automata in Computational Linguistics- LITP Spring School on Theoretical Computer Science : pp. 34-50.

Yvan Gut, Puteri Rashida Megat Ramli, Zaharin Yusoff, Chuah Choy Kim, Salina A. Samat, Christian Boitet, Nicolas Nédobejkine, Mathieu Lafourcade et al. (1996) *Kamus Perancis-Melayu Dewan, dictionnaire français-malais*. Dewan Bahasa Dan Pustaka, Kuala Lumpur, 667 p.

Mangeot-Lerebours, M. (1998). *Conception, implémentation et indexation de BaLeM, une base lexicale multilingue*. TALN'98, Paris, pp 215-218.

Mel'ãuk I. et al. (1984, 1988, 1992). *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques I, II et III*, Montréal(Quebec), Canada, Presses de l'université de Montréal, 172 p., 332p. et 323 p.

Mel'ãuk I., Clas A. & Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire.* Universités francophones et champs linguistiques, AUPELF-UREF et Duculot, Louvain la Neuve, 256 p.

Sérasset G. (1994a). *Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA*, COLING-94, 5-9 August 1994, vol. 1/2 : pp. 278-282.

Sérasset G. (1994b) *SUBLIM : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions*, Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1 : 194 p.

Sérasset G. (1997). *Le projet NADIA-DEC : vers un dictionnaire explicatif et combinatoire informatisé ?*, La mémoire des mots, 5ème journées scientifiques du réseau LTT, AUPELF-UREF, Tunis, à paraître.

Shieber S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar*, CSLI Lecture Notes, Center for the Study of Language and Information, Menlo Park, 105 p.